PSYCHOMETRIKA—VOL. 71, NO. 3, 457–461
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1501-4

MEASUREMENT WITHOUT COPPER INSTRUMENTS AND EXPERIMENT WITHOUT COMPLETE CONTROL

WILLEM J. HEISER
LEIDEN UNIVERSITY

1. Introduction

One basic reason that measurement in psychology requires statistics is that psychologists do not use copper instruments anymore, as they used to do in the nineteenth century. Instead, they determine test or total scores on the basis of miniature experiments with discrete outcomes, and use a variety of standard statististical techniques for reaching conclusions on the basis of observed data.[1] Borsboom (2006) wants us to believe that psychologists are seriously misled in their hope that they can make progress this way, and recommends an invasion of psychometricians carrying IRT missiles and SEM guns into psychology. My prediction is that such an invasion would simply be ignored. That is not to say that whenever psychometric modeling really makes a difference, no attempts should be made to reach the mainstream of psychology. Indeed, many psychometric contributions that are obsolete according to Borsboom, like Cronbach's alpha and exploratory factor analysis, in fact entered into the mainstream of psychology only because they tend to provide sensible answers to real problems, which cannot be easily surpassed. We should be more proud of them (even when a bit vulgarized), and carefully foster our accumulated knowledge base. Apart from strictly psychometric contributions, it has always been a task of psychometricians to introduce relevant new developments in the broad domain of mathematics and statistics into psychology. I am convinced that we should continue to do so, even when it concerns "observed score techniques" that are so detested in the focus article.

Borsboom is right in pointing out that the impact of IRT modeling in academic psychology is limited, and that problems of measurement invariance and test bias are ill-understood and neglected (but of course the IRT movement always had a primary focus on its major successes in another discipline, educational testing). Of the factors that he mentions as hindering the fruitful interplay between psychologists and psychometricians, I have no quarrel with the substantive and with most of the pragmatic ones, but I fail to see the relevance of the theoretical factors. In the following, I will try to explain why, and offer an important factor overlooked by Borsboom, which has to do with the changing relation between Cronbach's (1957) "two disciplines of scientific psychology."

2. Measurement without Copper Instruments

It is undoubtedly true that the single most important and typical contribution of psychometrics to both psychology and statistics is the latent variable. We have true scores, Thurstone's

Requests for reprints should be sent to Willem J. Heiser, Department of Psychology, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands. E-mail: Heiser@fsw.leideuniv.nl.

[1]The fact that psychologists so often have to deal with discrete outcomes (categorical data with only a few categories, ordinal data, counts), often with repeated measures, and sometimes with typically structured designs, defines the niche for psychometrics in the larger domain of statistics.

discriminal processes, factor analysis, multidimensional scaling models, structural equation models, item response models, and so on. Latent variables are hypothetical entities that may be fixed (parameters) or random (stochastic variables), and may also be classified in other ways. They are the basic elements of measurement models. This framework permits us to transform discrete psychological responses into numerical outcomes with known quality characteristics. Thus, the measurement model is a hypothetical copper instrument.

2.1. Role of the Latent Variable

It is often said that measurement models specify theoretical relations between concepts (latent variables) and observables. That sounds innocent enough, but as Borsboom convincingly shows in his example of measuring conscientiousness, it leads to an embarrassment of riches. I concur with Borsboom in that psychological theory is not strong enough to allow a motivated choice between specific psychometric models. But I also hold that this weakness is exactly why research psychologists use the sensible strategy to avoid reliance on hypothetical copper instruments whenever they can. As I have tried to argue elsewhere (Heiser, 2003), they are not interested in measurement per se, but in the establishment of (cause and effect) relations.

2.2. Attack Against Classical Test Theory

What can a psychologist do if she wants to report or use a concrete person score? It seems to me that calculation of scores unavoidably depends on the data. Under the condition of an embarrassment of riches, suppose she arbitrarily chooses to count those responses that are positive manifestations of the concept she wants to score. There are old results in psychometric theory that ensure that she cannot be too far off, compared to other scoring rules. Differences between weighted and unweighted counting tend to be small if the items being combined are correlated (Gullikson, 1950, p. 355). Also, under the same type of regularity conditions, there is a slightly nonlinear, but monotonic functional relation between a latent trait testimate and a true score estimate (Lord and Novick, 1968, p. 386), implying that the order of the person scores is identical. Recently, Warrens, De Gruijter, and Heiser (2006) showed that a similar relation exists between the latent trait person score and the optimal scaling person score.

It turns out that by far the most important consideration is that the items form a homogeneous set. There is again an embarrassment of riches in the choice of methods for finding a homogeneous set, but once found (approximately), different methods to calculate person scores are close to equivalent for research purposes (more care may be required for individual selection decisions). Classical test theory is not obsolete. Psychometric models at the item level are a refinement. Throwing classical test theory out of the window would only impair the credibility of psychometrics, and increase the gap with psychology.

2.3. Questionable Interpretations of Test Scores

The interpretation of principal components as "biologically based psychological tendencies," endowed with causal forces, as cited in Borsboom, is indeed a long stretch of the imagination. But one cannot blame principal components for this type of wishful thinking. Would it be any better with latent traits or factor scores? I do not think so. Perhaps we should put part of the blame on ourselves, since the idea that we can have causal models for correlated data without controlled interventions arose in our own field. Would it be possible that the language of variables with arrows pointing to each other in supposedly meaningful directions is giving the psychologist the false hope that he could discover causal relations instead of just relations?

2.4. Operationalism Rules

Borsboom is right that psychologists operationalize a lot, but that does not imply that they believe in philosophical operationalism. Neither do they need to believe in the psychometric dream of the hypothetical copper instrument, in which observables are related to theoretical attributes. Rather, like other scientists, psychologists tend to believe in the more general idea of approximation. As long as some protocol of data collection and/or method of data analysis can be justified, as providing an approximation of the psychological variable under study, it will do. Psychometricians might make fine distinctions between true scores and latent traits, or between formative and reflective models, for psychologists these are just two brands of approximations. They have to take a leap of faith anyhow, and it requires clear evidence of superiority in a variety of aspects for one particular method to become the preferred brand.

3. Experiment Without Complete Control

Apart from the measurement problem, there is a second reason why psychological research needs statistics: it is usually impossible to keep irrelevant variables fully under control. This difficulty is sometimes called *the third variable problem*. Picking up and ruling out third variables is the driving force of research design. Important methods of control are randomization (adding chance to the process!), factorial crossing, blocking, introducing covariates, and so forth. Although third variables can also be controlled ex post facto by regression methods, the Fisherian style of experimental thinking has caused a revolution in psychological research (Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989, Chap. 6). What are the consequences?

3.1. Samples Size Issues

Under this heading, Borsboom launches what at first appears to be a side attack against "experimentally oriented research" and the "various species of ANOVA," which would involve betting on observed score techniques and "stealing" assumptions. It is one thing that Borsboom cannot see the blessing of small sample statistics, but in any case, with this part of his argument he is not going to win the hearts of psychologists, who are discovering that the great thing of explanatory or independent variables is that you do not need to measure them.

3.2. The Rise of the Independent Variables

Borsboom underestimates the enormous impact of the movement toward more experimental research in psychology, with its emphasis on bringing situational variables under tight control, its attention for causal mechanisms by formulating "cause–effect hypotheses," and its tendency to regard individual differences as a nuisance since they increase within-treatment variance.[2] In contrast to what Borsboom believes, psychologists are much less frequently eye-balling correlations than they used to do. Fifty years ago, Cronbach (1957) could still write,

> In contrast to the Tight Little Island of the experimental discipline, correlational psychology is a sort of Holy Roman Empire whose citizens identify mainly with their own principalities. The discipline, the common service in which the principalities are united, is the study of correlations by Nature. While the experimenter is interested only in the variation he himself creates, the correlator finds his interest in the already

[2]Ironically, psychometricians specialized in IRT call the latent variable measuring individual differences a nuisance variable or, collectively, the nuisance parameters. Although the reason is different, the low esteem for the variations of Nature is equally disconcerting.

> existing variation between individuals, social groups, and species. (Cronbach, 1957, p. 672)

Cronbach also thought that "the tide of separation in psychology has already turned," and mentions as a prime example Meehl's introduction of construct validity in test theory, "capitalizing on the methodological and philosophical progress of the experimentalists" [*sic*!]. Nevertheless, I believe that fifty years later all signs are telling us that the experimentalists are winning big-time over the correlationists. The Holy Roman Empire is falling apart, and the Tight Little Island is growing into an archipelago where the sun never goes down.

The experimental method is triumphing in many areas outside traditional experimental psychology, like social, developmental, clinical, and even organizational psychology. Time and space do not permit going deeply into the reasons and effects of this revolution. But a major methodological aspect (and advantage for the psychologist) is that the independent variables need not be measured, but instead are manipulated, in which process they are reduced to attributes. If you study the effect of fear, there are many possible manipulations to create fear that can be compared with a control condition, but in any case the fear treatment is "on" or "off." Reduction to attributes implies that the theoretical model or reasoning to predict the outcome of the experiment can be qualitative instead of quantitative. Under the experimental method, psychology can reason in attributes, which explains what Borsboom calls "the almost complete absence of strong psychological theory." There is no need for quantification, except for the dependent variable.

3.3. Traits and States

Traits (either manifest or latent) can be dependent variables only in quasi–experiments, for instance when we compare monozygotic twins raised together and raised apart, since they are stable person characteristics. When the aim is to change the dependent variable by experimental manipulation, it must be a *state*. Quantitative state variables almost always involve counts, rates, ratings, or time, and are rather intricately related to the substantive research paradigm. They are never standardized with respect to some population, because effects of experimental manipulation are measured with respect to each other or with respect to a control group. They can also be more "quick and dirty" than standardized tests, because measurement error will only increase within-treatment variance, but not change between-treatment variance. Although effect size is negatively affected by measurement error, that involves a calculated risk and not a blind gamble; after all, a research paradigm comes into wider use only if its originator demonstrates that under typical circumstances of the setup reasonable to large effect sizes can be achieved.

4. Conclusion

Fisherian methodology rules in psychology, while the homeland of psychometrics is correlational methodology. Borsboom's timely discussion paper forces us to think hard about strategies that could save us from isolation and irrelevancy. Some of Borsboom's suggestions (write textbooks, publish widely) are fine. But from my analysis it should be clear that his suggestion to become an active psychological researcher is underestimating what it takes to be part of the psychological research community. For almost all of us, it would take an irreversible career change. When correlational psychology was still strong, one could imagine that psychometricians were a special kind of psychologist, since—after all—test theory was its infrastructure. But this is no longer the case.

There are signs that cognitive psychologists finally face individual differences as a serious factor and start modeling them. We should of course support this development whenever we can. It also appears that the mathematical psychology community finds itself in similar dangers as we do. I would strongly favor an attempt for rapprochement. It would add mass and focus if we had a united platform for the whole of quantitative psychology, following the motto on the cover of this journal. In the recent past, some of our colleagues have made a career change to statistics, but such a move only rarely increased their impact. Psychometrics is a discipline by itself, with a body of results that stood the test of time, but now it has to find a new balance between psychology and statistics.

Latent variables are important, but we should not try to push them at all costs, and they can no longer be the dominant instrument in our repertoire. Psychology needs a new generation of statistical techniques adapted to its current challenges. Physiological outcome measures are hot, so there is a need for functional data analysis. More hierarchical data are collected, so there is a need for multilevel analysis. There is increased interest in moderator variables, so we should work on regression trees in (quasi-) experimental setups. These are just a few examples to broaden the scope of psychometrics.

Finally, psychometrics should care more about its image in the outside world. What we do not do enough of—and I blame myself, too—is propagating and defending our heritage in the larger scientific and public community. We should follow the example of people like John Carroll, who took a brave stand against Stephen Jay Gould's biased views on mental testing and factor analysis (Carroll, 1995). Join forces with friends, and attack that enemy!

References

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440.
Carroll, J.B. (1995). Reflections on Stephen Jay Gould's *The mismeasure of man* (1981). *Intelligence, 21*, 121–134.
Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance*. Cambridge: Cambridge University Press.
Gullikson, H. (1950). *Theory of mental tests*. New York: Wiley.
Heiser, W.J. (2003). Trust in relations. *Measurement: Interdisciplinary Research and Perspectives, 1*, 264–269.
Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Warrens, M.J., De Gruijter, D.N.M., & Heiser, W.J. (2006). A systematic comparison between classical optimal scaling and the two-parameter IRT model. *Applied Psychological Measurement*, *30*, 1–15.

Manuscript received 30 JUN 2006
Final version received 30 JUN 2006
Published Online Date: 23 SEP 2006